BUSINESS STATISTICS COMM 215
Lecturer: Samie Li Shang Ly
Lesson 10/11: Simple Linear Regression

PROBLEM # 10.2

The linear regression model requires the following assumptions:

a. For any given value of x, the y values are normally distributed with a mean that is on the regression line.

b. Regardless of the value of x, the standard deviation of the distribution of y values about the regression line is constant; this is the assumption of homoscedasticity.

c. Each value of y is statistically independent from all the other values of y.

These assumptions can be restated in terms of the error term:

a. For any given value of x, the error terms will be normally distributed with a mean of 0.

b. The standard deviation for the distribution of error terms is constant regardless of the value of x.

c. The values of the error term are statistically independent.

PROBLEM # 10.6

The tables below show the sum of squares value for each line.

Let $\hat{y} = 5 + 3x$:

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
3	8	14	-6	36
5	18	20	-2	4
7	30	26	4	16
9	32	32	0	0
				sum = 56

Let $\hat{y} = -2 + 4x$:

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
3	8	10	-2	4
5	18	18	0	0
7	30	26	4	16
9	32	34	-2	4

				sum = 24

Using the least squares criterion, the second line fits the data better.

PROBLEM # 10.8

a.

This exercise can be solved using a pocket calculator and the method shown in the solution to exercise 15.9. We will use Minitab. In generating the printout shown below, we have specified that a prediction be made for trailer sales when boat sales = 500 thousand.

Regression Analysis: Trailers versus Boats

The regression equation is Trailers = - 166 + 0.577 Boats

Predictor	Coef	SE Coef	T	P
Constant	-165.67	52.15	-3.18	0.034
Boats	0.57711	0.08686	6.64	0.003

S = 5.66591 R-Sq = 91.7% R-Sq(adj) = 89.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1417.1	1417.1	44.14	0.003
Residual Error	4	128.4	32.1		
Total	5	1545.5			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	122.89	8.97	(97.97, 147.80)	(93.42, 152.35)XX

XX denotes a point that is an extreme outlier in the predictors.

Values of Predictors for New Observations

New Obs	Boats

```
1      500
```

a. To the greatest number of decimal places in the printout, the regression equation is

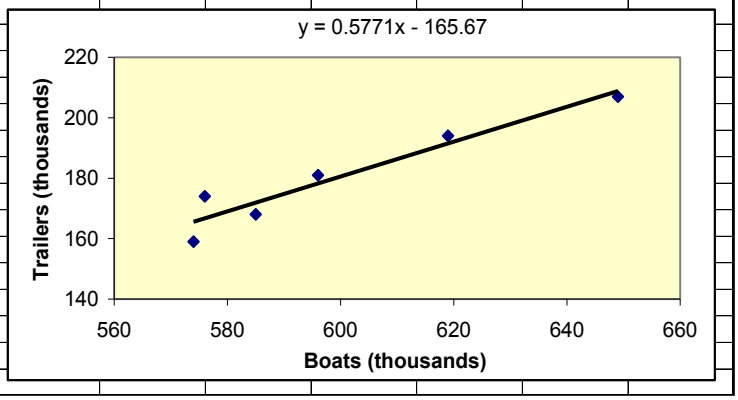
Trailers = -165.67 + 0.57711*Boats. Since the slope is positive, we can deduce that boat trailer sales increase as boat sales increase. Moreover, the value of the slope implies that for every additional thousand boats sold, there are 0.57711 additional thousands of trailers sold.

b. Substituting Boats = 500 thousand into the equation in part a, the estimated value of Trailers will be 122.89 thousand, as shown in the "Fit" column of the printout.

c. Perhaps boat owners do not buy a new trailer every time they buy a new boat. Alternatively, boat owners might make their own trailers or they may haul their boats in the back of a truck or on the roof of their car or truck.

Applying the procedure described in textbook chapter 2, Computer Solutions 2.7, we obtain the corresponding Excel plot and equation shown below.

	A	B	C	D	E	F	G	H	I
1	Boats	Trailers							
2	649	207							
3	619	194							
4	596	181							
5	576	174							
6	585	168							
7	574	159							
8									
9									
10									
11									
12									
13									
14									
15									



PROBLEM # 10.16

a.

Age	Repairs
110	327.67
113	376.68
114	392.52
134	443.14
93	342.62
141	476.16
115	324.74
115	338.98
115	433.45
142	526.37
96	362.42
139	448.76
89	335.27
93	350.94
91	291.81
109	467.80

138	474.48
83	354.15
100	420.11
137	416.04

a $b_1 = \frac{s_{xy}}{s_x^2} = \frac{936.82}{378.77} = 2.47$ $b_0 = \bar{y} - b_1\bar{x} = 395.21 - 2.47(113.35) = 115.24.$

Regression line: $\hat{y} = 115.24 + 2.47x$ (Excel: $\hat{y} = 114.85 + 2.47x$)

b $b_1 = 2.47$; for each additional month of age, repair costs increase on average by $2.47.

$b_0 = 114.85$ is the y-intercept.

c $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(936.82)^2}{(378.77)(4{,}094.79)} = .5659$ (Excel: $R^2 = .5659$) 56.59% of the variation in repair costs s explained by the variation in ages.

d $SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = (20-1)\left(4{,}094.79 - \frac{(936.82)^2}{378.77}\right) = 33{,}777$

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{33{,}777}{20-2}} = 43.32$ (Excel: $s_\varepsilon = 43.32$).

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Rejection region: $t > t_{\alpha/2,n-2} = t_{.025,18} = 2.101$ or $t < -t_{\alpha/2,n-2} = -t_{.025,18} = -2.101$

$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} = \frac{43.32}{\sqrt{(20-1)(378.77)}} = .511$

$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{2.47 - 0}{.511} = 4.84$ (Excel: t = 4.84, p–value = .0001. There is enough evidence to infer that repair costs and age are linearly related.

e $\hat{y} = b_0 + b_1 x_g = 115.24 + 2.47(120) = 411.64$

Prediction interval: $\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{(n-1)s_x^2}}$ (where $t_{\alpha/2,n-2} = t_{.025,18} = 2.101$)

$$= 411.64 \pm 2.101(43.32)\sqrt{1 + \frac{1}{20} + \frac{(120 - 113.35)^2}{(20-1)(378.77)}} = 411.64 \pm 93.54$$

Lower prediction limit = 318.1, upper prediction limit = 505.2 (Excel: 318.1, 505.2)

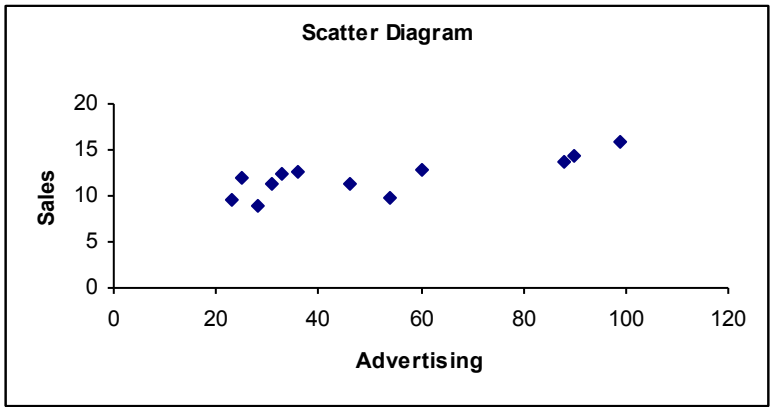
PROBLEM # 10.17

The prediction interval for y gets wider as the x value on which the interval estimate is based gets farther away from the mean of x because there is less error in making interval estimates based on x values that are closer to the mean. This can be seen in the formula; the numerator of the fraction under the square root includes $(x - \bar{x})^2$. This value, of course, increases as the x value moves away from the mean.

PROBLEM # 10.21

The coefficient of correlation (r) describes both the direction and the strength of the linear relationship between two variables. The coefficient of determination (r^2) expresses the proportion of the variation in the dependent variable (y) that is explained by the regression line, $\hat{y} = b_0 + b_1 x$, but it does not indicate the direction of the relationship.

PROBLEM # 10.22



b	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
23	9.6	529	92.16	220.8	
46	11.3	2,116	127.69	519.8	
60	12.8	3,600	163.84	768.0	
54	9.8	2,916	96.04	529.2	

28	8.9	784	79.21	249.2
33	12.5	1,089	156.25	412.5
25	12.0	625	144.00	300.0
31	11.4	961	129.96	353.4
36	12.6	1,296	158.76	453.6
88	13.7	7,744	187.69	1205.6
90	14.4	8,100	207.36	1296.0
99	15.9	9,801	252.81	1,574.1
Total	613	144.9	39,561	1,795.77
7,882.2				

$$\sum_{i=1}^{n} x_i = 613 \qquad \sum_{i=1}^{n} y_i = 144.9 \qquad \sum_{i=1}^{n} x_i^2 = 39{,}561 \qquad \sum_{i=1}^{n} x_i y_i = 7{,}882.2$$

$$s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right] = \frac{1}{12-1}\left[7{,}882.2 - \frac{(613)(144.9}{12}\right] = 43.66$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{12-1}\left[39{,}561 - \frac{(613)^2}{12}\right] = 749.7$$

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{43.66}{749.7} = .0582$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{613}{12} = 51.08$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{144.9}{12} = 12.08$$

$b_0 = \bar{y} - b_1\bar{x} = 12.08 - (.0582)(51.08) = 9.107$

The sample regression line is

$\hat{y} = 9.107 + .0582x$

The slope tells us that for each additional thousand dollars of advertising sales increase on average by .0582 million. The y-intercept has no practical meaning.

PROBLEM # 10.23

Determine the standard error of estimate

a. Is there evidence of a linear relationship between advertising and sales?
b. Estimate β_1 with 95% confidence.
c. Compute the coefficient of determination and interpret the value.
d. Briefly summarize what you have learned in parts c,d,e, and f.

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] = \frac{1}{12-1}\left[1{,}795.77 - \frac{(144.9)^2}{12}\right] = 4.191$$

$$SSE = (n-1)\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right) = (12-1)\left(4.191 - \frac{(43.66)^2}{749.7}\right) = 18.13$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{18.13}{12-2}} = 1.347 \text{ (Excel: } s_\varepsilon = 1.347)$$

d $H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Rejection region: $t > t_{\alpha/2,n-2} = t_{.025,10} = 2.228$ or $t < -t_{\alpha/2,n-2} = -t_{.025,10} = -2.228$

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} = \frac{1.347}{\sqrt{(12-1)(749.7)}} = .0148$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{.0582 - 0}{.0148} = 3.93$$ (Excel: t = 3.93, p–value = .0028. There is enough evidence to infer a linear relationship between advertising and sales.

e $b_1 \pm t_{\alpha/2,n-2}s_{b_1} = .0582 \pm 2.228(.0148) = .0582 \pm .0330$ LCL = .0252, UCL = .0912

f $R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(43.66)^2}{(749.7)(4.191)} = .6067$ (Excel: $R^2 = .6066$). 60.67% of the variation in sales is explained by the variation in advertising.

g There is evidence of a linear relationship. For each additional dollar of advertising sales increase, on average by .0582.

PROBLEM # 10.24

We know that the coefficient of determination is $r^2 = 1 - (SSE/SST)$. Therefore the coefficient of determination is 1 - (40.0/200.0) = 0.80.

PROBLEM # 10.25

a. The monthly mortgage payment would likely be directly related to the market value of the house, the interest rate, the size of the house, or the monthly taxes and insurance, among other variables.

b. The monthly mortgage payment would likely be inversely related to the age of the house, among other variables.

c. The monthly mortgage payment would likely be unrelated to the amount of chocolate consumed by the owners, and a wide variety of other variables.

PROBLEM # 10.26- A tire company has carried out tests in which rolling resistance (pounds) and inflation pressure (pounds per square inch, or psi) have been measured for psi values ranging from 20 to 45. The regression analysis is summarized in the following MiniTab printout:

Regression Analysis

The regression equation is

ROLRESIS = 9.45 – 0.0811 PSI

Predictor	Coef	StDev	T	P
Constant	9.450	1.228	7.69	0.000
PSI	-0.08113	0.03416	-2.38	0.029

S= 0.8808 R-Sq = 23.9% R-Sq (adj) = 19.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4.3766	4.3766	5.64	0.029
Error	18	13.9657	0.7759		
Total	19	18.3422			

a. The least squares regression line is: ROLRESIS = 9.450 - 0.08113 PSI.

b. 23.9% of the variation in rolling resistance is explained by the regression line. (See R-sq.)

c. The slope of the line differs from zero at the 0.029 level of significance. Minitab used a two-tail t-test to reach this conclusion.

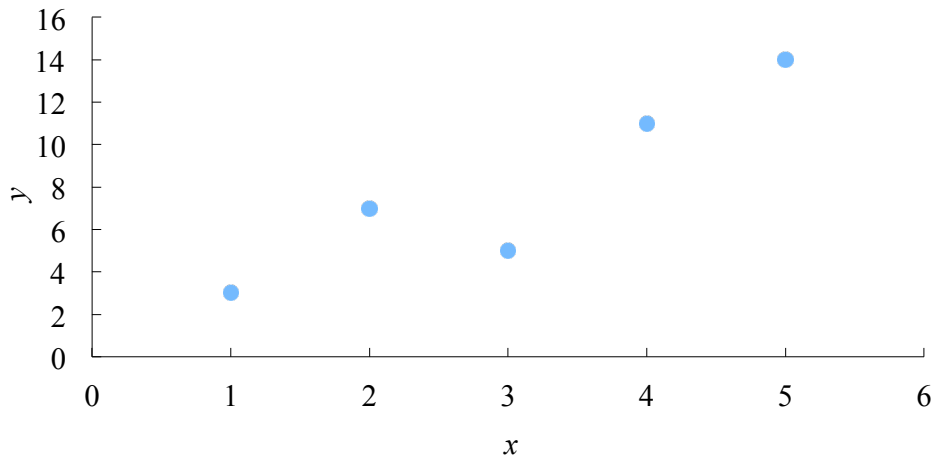
d. The coefficient of correlation differs from zero at the 0.029 level of significance. (See the ANOVA table.) This is the same level found in part c. These levels will always be the same because the tests are equivalent.

e. The 95% confidence interval for the slope of the population regression line is:

$b_1 \pm t\, s_{b_1}$ = -0.08113 ± 2.101(0.03416) = -0.08113 ± 0.07177, or from -0.15290 to -0.00936.

Least Square Method: Suggested: page 569 # 1,3,4,6 Webfile Honda Accord ,9 Webfile Suitcases, 14 Webfile Laptop

1 a.



b. There appears to be a positive linear relationship between *x* and y.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between *x and* y; in part (d) we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{15}{5} = 3 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{40}{5} = 8$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 26 \quad \Sigma(x_i - \bar{x})^2 = 10$

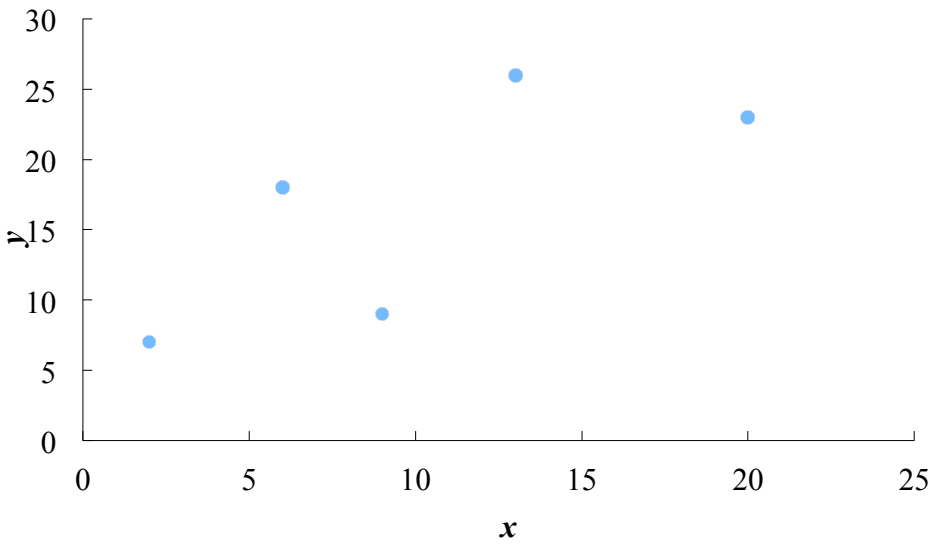
$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{26}{10} = 2.6$

$b_0 = \bar{y} - b_1\bar{x} = 8 - (2.6)(3) = 0.2$

$\hat{y} = 0.2 + 2.6x$

e. $\hat{y} = 0.2 + 2.6(4) = 10.6$

3. a.



b. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{50}{5} = 10 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{83}{5} = 16.6$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 171 \quad \Sigma(x_i - \bar{x})^2 = 190$

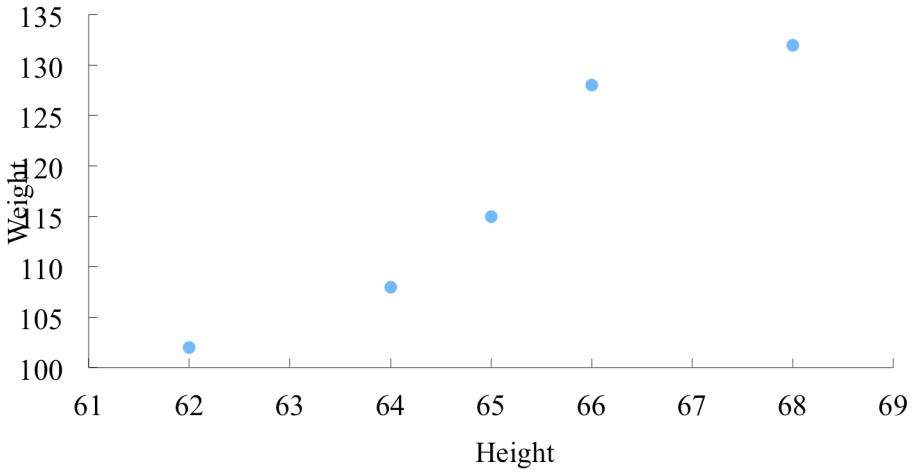
$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{171}{190} = 0.9$

$b_0 = \bar{y} - b_1\bar{x} = 16.6 - (0.9)(10) = 7.6$

$\hat{y} = 7.6 + 0.9x$

c. $\hat{y} = 7.6 + 0.9(6) = 13$

4. a.



b. There appears to be a positive linear relationship between x = height and y = weight.

c. Many different straight lines can be drawn to provide a linear approximation of the relationship between x *and* y; in part (d) we will determine the equation of a straight line that "best" represents the relationship according to the least squares criterion.

d. $\bar{x} = \dfrac{\Sigma x_i}{n} = \dfrac{325}{5} = 65 \quad \bar{y} = \dfrac{\Sigma y_i}{n} = \dfrac{585}{5} = 117$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 110 \quad \Sigma(x_i - \bar{x})^2 = 20$

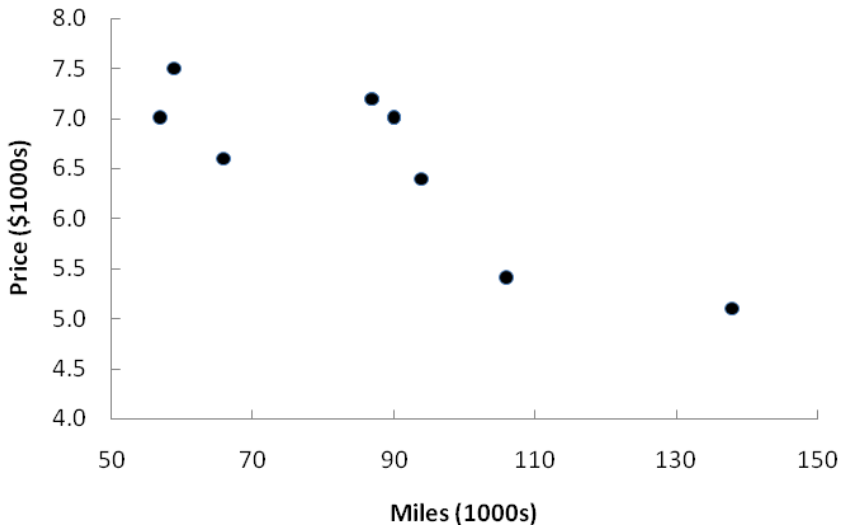
$b_1 = \dfrac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \dfrac{110}{20} = 5.5$

$b_0 = \bar{y} - b_1\bar{x} = 117 - (5.5)(65) = -240.5$

$\hat{y} = -240.5 + 5.5x$

e. $\hat{y} = -240.5 + 5.5x = -240.5 + 5.5(63) = 106$ pounds

6. a.



b. There appears to be a negative linear relationship between x = miles and y = sales price.

If the car has higher miles, the sales price tends to be lower.

c. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{874}{10} = 87.4 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{66.4}{10} = 6.64$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = -135.66 \quad \Sigma(x_i - \bar{x})^2 = 5152.4$

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{-135.66}{5152.40} = -.02633$

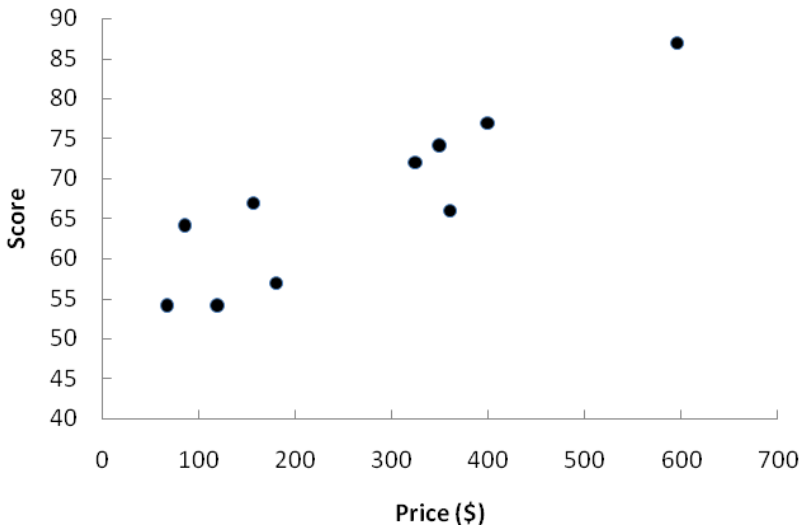
$b_0 = \bar{y} - b_1\bar{x} = 6.64 - (-.02633)(87.4) = 8.9412$

$\hat{y} = 8.9412 - .02633x$

d. The slope of the estimated regression equation is -.02633. Thus, a one unit increase in the value of x will result in a decrease in the estimated value of y equal to .02633. Because the data were recorded in thousands, every additional 1000 miles on the car's odometer will result in a $26.33 decrease in the estimated price.

e. $\hat{y} = 8.9412 - .02633(100) = 6.3 \text{ or } \$6300$

9. a.



b. There appears to be a positive linear relationship between *x* = price and *y* = score.

c. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{2638}{10} = 263.8 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{672}{10} = 67.2$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 14,601.40 \quad \Sigma(x_i - \bar{x})^2 = 258,695.60$

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{14,601.40}{258,695.60} = .05644$

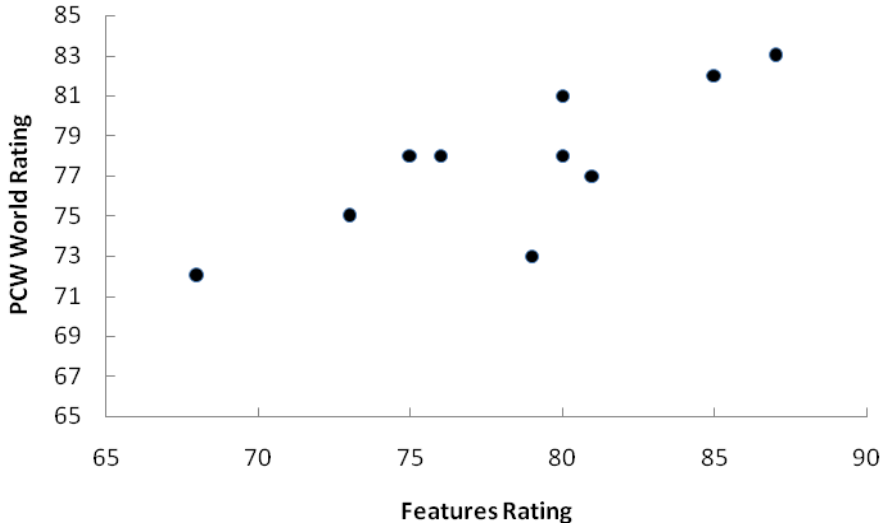
$b_0 = \bar{y} - b_1\bar{x} = 67.2 - (.05644)(263.8) = 52.311$

$\hat{y} = 52.311 + .05644x$

d. The slope is .05644. For a $100 higher price, the score can be expected to increase 100(.05644) = 5.644, or about 6 points.

e. $\hat{y} = 52.311 + .05644(225) = 65$

14. a.



b. There appears to be a positive linear relationship between x = features rating and y = PCW World Rating.

c. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{784}{10} = 78.4 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{777}{10} = 77.7$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 147.20 \quad \Sigma(x_i - \bar{x})^2 = 284.40$

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{147.20}{284.40} = .51758$

$b_0 = \bar{y} - b_1\bar{x} = 77.7 - (.51758)(78.4) = 37.1217$

$\hat{y} = 37.1217 + .51758x$

d. $\hat{y} = 37.1217 + .51758(70) = 73.35$ or 73

Coefficient of Determination: Suggested: page 581 # 15,17,18,19 Webfile Sales,21

15. a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = 0.2 + 2.6x_i \qquad \bar{y} = 8$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 12.40 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 80$

Thus, SSR = SST - SSE = 80 - 12.4 = 67.6

b. r^2 = SSR/SST = 67.6/80 = .845

The least squares line provided a very good fit; 84.5% of the variability in y has been explained by the least squares line.

c. $r_{xy} = \sqrt{.845} = +.9192$

17. The estimated regression equation and the mean for the dependent variable are:

$\hat{y}_i = 7.6 + .9x \qquad \bar{y} = 16.6$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 127.3 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 281.2$

Thus, SSR = SST - SSE = 281.2 – 127.3 = 153.9

r^2 = SSR/SST = 153.9/281.2 = .547

We see that 54.7% of the variability in *y* has been explained by the least squares line.

$r_{xy} = \sqrt{.547} = +.740$

18. a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y} = 1790.5 + 581.1x \qquad \bar{y} = 3650$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 85{,}135.14 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 335{,}000$

Thus, SSR = SST - SSE = 335,000 - 85,135.14 = 249,864.86

b. r^2 = SSR/SST = 249,864.86/335,000 = .746

We see that 74.6% of the variability in *y* has been explained by the least squares line.

c. $r_{xy} = \sqrt{.746} = +.8637$

19. a. The estimated regression equation and the mean for the dependent variable are:

$\hat{y} = 80 + 4x \qquad \bar{y} = 108$

The sum of squares due to error and the total sum of squares are

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 170 \qquad \text{SST} = \sum(y_i - \bar{y})^2 = 2442$

Thus, SSR = SST - SSE = 2442 - 170 = 2272

b. r^2 = SSR/SST = 2272/2442 = .93

We see that 93% of the variability in *y* has been explained by the least squares line.

c. $r_{xy} = \sqrt{.93} = +.96$

21. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{3450}{6} = 575 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{33,700}{6} = 5616.67$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 712,500 \quad \Sigma(x_i - \bar{x})^2 = 93,750$

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{712,500}{93,750} = 7.6$

$b_0 = \bar{y} - b_1\bar{x} = 5616.67 - (7.6)(575) = 1246.67$

$\hat{y} = 1246.67 + 7.6x$

b. $7.60

c. The sum of squares due to error and the total sum of squares are:

$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = 233,333.33 \quad \text{SST} = \sum(y_i - \bar{y})^2 = 5,648,333.33$

Thus, SSR = SST - SSE = 5,648,333.33 - 233,333.33 = 5,415,000

r^2 = SSR/SST = 5,415,000/5,648,333.33 = .9587

We see that 95.87% of the variability in *y* has been explained by the estimated regression equation.

d. $\hat{y} = 1246.67 + 7.6x = 1246.67 + 7.6(500) = \$5046.67$

Testing for Significance: Suggested: page 592 # 23,25,27 Webfile Boots

23. a. s^2 = MSE = SSE / (*n* - 2) = 12.4 / 3 = 4.133

b. $s = \sqrt{\text{MSE}} = \sqrt{4.133} = 2.033$

c. $\Sigma(x_i - \bar{x})^2 = 10$

$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{2.033}{\sqrt{10}} = 0.643$

d. $t = \frac{b_1}{s_{b_1}} = \frac{2.6}{.643} = 4.044$

Using *t* table (3 degrees of freedom), area in tail is between .01 and .025

p-value is between .02 and .05

Using Excel or Minitab, the *p*-value corresponding to *t* = 4.04 is .0272.

Because *p*-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

e. MSR = SSR / 1 = 67.6

F = MSR / MSE = 67.6 / 4.133 = 16.36

Using F table (1 degree of freedom numerator and 3 denominator), p-value is between .025 and .05

Using Excel or Minitab, the p-value corresponding to F = 16.36 is .0272.

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p-value
Regression	67.6	1	67.6	16.36	.0272
Error	12.4	3	4.133		
Total	80.0	4			

25. a. s^2 = MSE = SSE/(n - 2) = 127.3/3 = 42.4333

$$s = \sqrt{\text{MSE}} = \sqrt{42.4333} = 6.5141$$

b. $\Sigma(x_i - \bar{x})^2 = 190$

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{6.5141}{\sqrt{190}} = 0.4726$$

$$t = \frac{b_1}{s_{b_1}} = \frac{.9}{.4726} = 1.90$$

Using t table (3 degrees of freedom), area in tail is between .05 and .10

p-value is between .10 and .20

Using Excel or Minitab, the p-value corresponding to t = 1.90 is .1530.

Because p-value $> \alpha$, we cannot reject H_0: $\beta_1 = 0$; x and y do not appear to be related.

c. MSR = SSR/1 = 153.9 /1 = 153.9

F = MSR/MSE = 153.9/42.4333 = 3.63

Using F table (1 degree of freedom numerator and 3 denominator), p-value is greater than .10

Using Excel or Minitab, the p-value corresponding to F = 3.63 is .1530.

Because p-value $> \alpha$, we cannot reject H_0: $\beta_1 = 0$; x and y do not appear to be related.

27. a. $\bar{x} = \frac{\Sigma x_i}{n} = \frac{37}{10} = 3.7 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{1654}{10} = 165.4$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 315.2 \quad \Sigma(x_i - \bar{x})^2 = 10.1$

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{315.2}{10.1} = 31.2079$$

$$b_0 = \bar{y} - b_1\bar{x} = 165.4 - (31.2079)(3.7) = 49.9308$$

$$\hat{y} = 49.9308 + 31.2079x$$

b. SSE = $\Sigma(y_i - \hat{y}_i)^2 = 2487.66$ SST = $\Sigma(y_i - \bar{y})^2 = 12{,}324.4$

Thus, SSR = SST - SSE = 12,324.4 - 2487.66 = 9836.74

MSR = SSR/1 = 9836.74

MSE = SSE/(n - 2) = 2487.66/8 = 310.96

F = MSR / MSE = 9836.74/310.96 = 31.63

Using F table (1 degree of freedom numerator and 8 denominator), p-value is less than .01

Using Excel or Minitab, the p-value corresponding to F = 31.63 is .001.

Because p-value $\leq \alpha$, we reject H_0: $\beta_1 = 0$

Upper support and price are related.

c. r^2 = SSR/SST = 9,836.74/12,324.4 = .80

The estimated regression equation provided a good fit; we should feel comfortable using the estimated regression equation to estimate the price given the upper support rating.

d. $\hat{y}$ = 49.93 + 31.21(4) = 174.77

Using the estimated regression equation for estimation and prediction: Suggested: page 599, # 33, 37, 39

33. a. s = 8.7560

b. $\bar{x} = 11 \quad \Sigma(x_i - \bar{x})^2 = 180$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 8.7560\sqrt{\frac{1}{5} + \frac{(8-11)^2}{180}} = 4.3780$$

$$\hat{y} = 68 - 3x = 68 - 3(8) = 44$$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$44 \pm 3.182\ (4.3780) = 44 \pm 13.93$

or 30.07 to 57.93

c. $s_{\text{ind}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 8.7560\sqrt{1 + \frac{1}{5} + \frac{(8-11)^2}{180}} = 9.7895$

d. $\hat{y}_p \pm t_{\alpha/2} s_{\text{ind}}$

$44 \pm 3.182(9.7895) = 44 \pm 31.15$

or 12.85 to 75.15

37. a. $\bar{x} = 57 \quad \Sigma(x_i - \bar{x})^2 = 7648$

$s^2 = 1.88 \quad s = 1.37$

$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 1.37\sqrt{\frac{1}{7} + \frac{(52.5-57)^2}{7648}} = 0.52$

$\hat{y}_p \pm t_{\alpha/2} s_{\hat{y}_p}$

$\hat{y} = 4.68 + 0.16x = 4.68 + 0.16(52.5) = 13.08$

$13.08 \pm 2.571\ (.52) = 13.08 \pm 1.34$

or 11.74 to 14.42 or \$11,740 to \$14,420

b. $s_{\text{ind}} = 1.47$

$13.08 \pm 2.571\ (1.47) = 13.08 \pm 3.78$

or 9.30 to 16.86 or \$9,300 to \$16,860

c. Yes, \$20,400 is much larger than anticipated.

d. Any deductions exceeding the \$16,860 upper limit could suggest an audit.

39. a. Let x = miles of track and y = weekday ridership in thousands.

$\bar{x} = \frac{\Sigma x_i}{n} = \frac{203}{7} = 29 \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{309}{7} = 44.1429$

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 1471 \quad \Sigma(x_i - \bar{x})^2 = 838$

$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} = \frac{1471}{838} = 1.7554$

$b_0 = \bar{y} - b_1\bar{x} = 44.1429 - (1.7554)(29) = -6.76$

$\hat{y} = -6.76 + 1.755x$

b. SST =3620.9 SSE = 1038.7 SSR = 2582.1

r^2 = SSR/SST = 2582.1/3620.9 = .713

The estimated regression equation explained 71.3% of the variability in *y*; a good fit.

c. s^2 = MSE = 1038.7/5 = 207.7

$s = \sqrt{207.7} = 14.41$

$$s_{\hat{y}_p} = s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 14.41\sqrt{\frac{1}{7} + \frac{(30-29)^2}{838}} = 5.47$$

$\hat{y} = -6.76 + 1.755x = -6.76 + 1.755(30) = 45.9$

$\hat{y}_p \pm t_{\alpha/2}s_{\hat{y}_p}$

45.9 ± 2.571(5.47) = 45.9 ± 14.1

or 31.8 to 60

d. $$s_{ind} = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}} = 14.41\sqrt{1 + \frac{1}{7} + \frac{(30-29)^2}{838}} = 15.41$$

$\hat{y}_p \pm t_{\alpha/2}s_{ind}$

45.9 ± 2.571(15.41) = 45.9 ± 39.6

or 6.3 to 85.5

The prediction interval is so wide that it would not be of much value in the planning process. A larger data set would be beneficial.

End of chapter problems page 625 # 55,57,58 Webfile IPO,60 Webfile Online Edu

55. No. Regression or correlation analysis can never prove that two variables are causally related.

57. The purpose of testing whether $\beta_1 = 0$ is to determine whether or not there is a significant relationship between *x* and *y*. However, rejecting $\beta_1 = 0$ does not necessarily imply a good fit. For example, if $\beta_1 = 0$ is rejected and r^2 is low, there is a statistically significant relationship between *x* and *y* but the fit is not very good.

58. a. The Minitab output is shown below:

```
The regression equation is
```

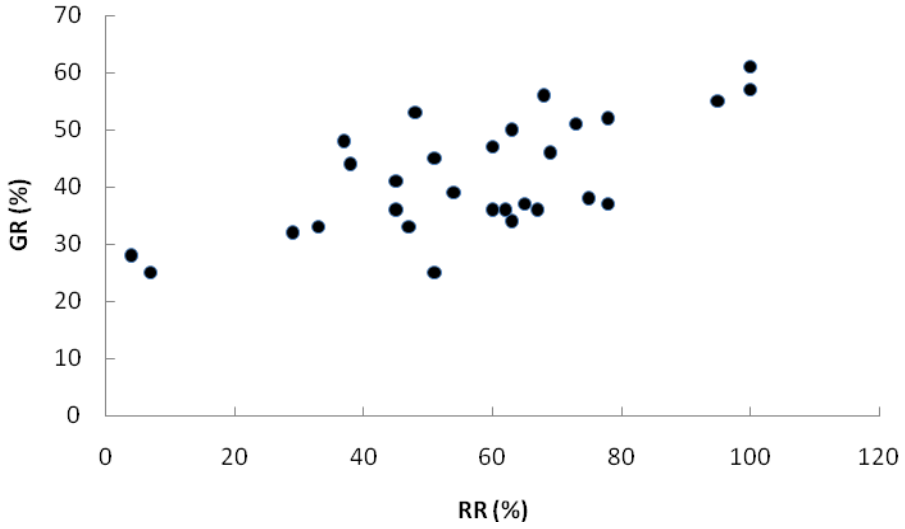
```
Price = 9.26 + 0.711 Shares

Predictor        Coef     SE Coef          T        P
Constant        9.265       1.099       8.43    0.000
Shares         0.7105      0.1474       4.82    0.001

S = 1.419       R-Sq = 74.4%     R-Sq(adj) = 71.2%

Analysis of Variance

Source             DF          SS          MS          F
P
Regression          1      46.784      46.784      23.22
0.001
Residual Error      8      16.116       2.015
Total               9      62.900
```

b. Since the *p*-value corresponding to F = 23.22 = .001 < α = .05, the relationship is significant.

c. r^2 = .744; a good fit. The least squares line explained 74.4% of the variability in Price.

d. $\hat{y} = 9.26 + .711(6) = 13.53$

60. a.



The scatter diagram indicates a positive linear relationship between the two variables. Online universities with higher retention rates tend to have higher graduation rates.